The name of the disease I have selected is Fibrodysplasia ossificans progressiva. It is a genetic disease which causes the muscle tissue and connective tissue of a sufferer to be replaced by bone, generally after trauma (even very mild) to the body part. The human name for the gene that is thought to be involved is activin A receptor type 1 and the official symbol is ACVR1. This gene is known by other names, namely activin A receptor type I, activin A receptor, type I, activin A receptor, type II-like kinase 2, activin A type I receptor, activin A type I receptor precursor, ActR-IA protein, human, ACTRI, ACVR1\_HUMAN, ACVR1A, ACVRLK2, ALK2, hydroxyalkyl-protein kinase, SKR1, according to the US national library of medicine. NCBI also lists FOP and TSRI as additional names for this gene.

Is there a homologue present in a model system such as the mouse, the fruit fly or yeast? If so, give reasons why you think this is a true homologue; if not, explain what you did to try to find one.

There are several homologues present in other species. P.troglodytes, M.mulatta, C.lupus, B.taurus, M.musculus, R.norvegicus, G.gallus, X.tropicalis, D.rerio, D.melanogaster and A.gambiae.

NCBI HomoloGene

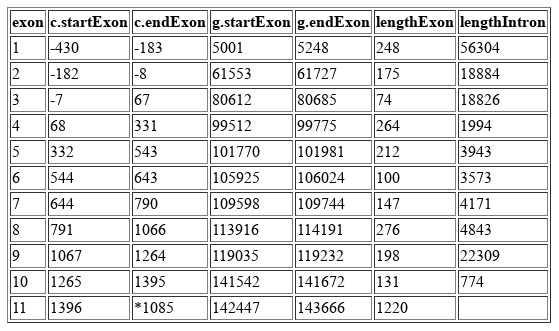
**Question 2. (3 marks)**  
Now investigate the structure of the gene. Find the gene in a database. Which chromosome is it located on, and at which position along the chromosome?

Chromosome 2

The NCBI lists the location for the gene as 2q24.1. This means that the gene is located on the second chromosome, at region 2, band 4, sub-band 1 of the longer arm.

What is the structure of the gene, how long is it, how many exons and introns does it have, and what percentages are exon and intron?

The gene has 16 exons and 15 introns [genomic context NCBI]. The [NM\_001105.4](https://databases.lovd.nl/shared/transcripts/00002047) transcript, however, has 11 exons and 10 introns [<https://databases.lovd.nl/shared/refseq/ACVR1_NM_001105.4_table.html> Leiden open Variation databse].



This is the table for a transcript of [**NM\_001105.4**](https://www.ncbi.nlm.nih.gov/nuccore/NM_001105.4) **'s** exons and introns taken from the Leiden open variation database.

The percentage of exons in the transcript is (3045/138666)\*100 = ~2.20% and the percentage of introns is (135,621/138666)\*100 = ~97.8%.

Download the gene transcript and the coding sequences for your gene - if multiple are listed, choose the main one, at the top of the list. Where did you get this sequence from and what was the unique identifier used so that someone else could be sure they were looking at the same sequence?

I go the transcript and coding sequences from the NCBI website, by searching for ACVR1 in the gene database, scrolling down the page to the [**NCBI Reference Sequences (RefSeq)**](https://www.ncbi.nlm.nih.gov/gene/90) section, clicking on the top entry [**NM\_001105.4**](https://www.ncbi.nlm.nih.gov/nuccore/NM_001105.4) as downloading both sequences as FASTA files. The unique identifier used so that someone else could be sure they were looking at the same sequences was [**NM\_001105.4**](https://www.ncbi.nlm.nih.gov/nuccore/NM_001105.4) .

How long is the transcript, and what proportion is coding?

The length of the transcript is 2300 and the length of the coding sequence is 1584, which means that the proportion that is coding is around 68.9%.

**Question 3. (3 marks)**Translate your cDNA sequence into protein/amino acid sequence. How many amino acids does your protein contain?

My protein contains all 20 amino acids and 61 out of 64 possible codons. The most common amino acid in the protein is Leucine (L). Six codons for this amino acid exist. The following table illustrates how often each is used:

|  |  |
| --- | --- |
| Codon | Times used in my protein |
| CTT | 10 |
| CTC | 7 |
| CTA | 14 |
| CTG | 2 |
| TTA | 10 |
| TTG | 10 |

Information in the table found using http://www.bioinformatics.org/sms2/codon\_usage.html

The codon usage database lists the frequency which each codon is used in a species (different species prefer different codons). Sequences which have too many rarer codons result in slowing down transcription and inhibition of protein expression - in extreme cases, rare codons are thought to introduce transcription errors when the rare tRNA is not available. There is no hard threshold, but generally codons with 1% usage or less are considered rare. [from assignment sheet]

The codon frequency table is given in frequency/thousand. Therefore, to find codons with 1% usage or less, I need to search for the codons which have a frequency of 10/thousand are less. These codons are: UCG, UAA no, UAG no, UGU, UGC, UGA, CUC, CCC, CCG, CAC, CGU, CGC, CGA, CGG, ACG, AGC, AGG, GCG no, GGC and GGG. These all occur in my protein except for UAA, UAG (both stop codons) and GCG. Therefore, all except those three may cause problems if I were to try and express my human cDNA in yeast.

1. CTGAAGCGGGAGGCTGAGACGCTGCGGGAGCGGGAGGGC

Mus musculus house mouse Kcnb1

potassium voltage gated channel, Shab-related subfamily, member 1  
2. CTCAAGCGTGAGGCCGAGACCCTACGGGAGCGGGAAGGC

human KCNB1 potassium voltage-gated channel subfamily B member 1  
3. GAAGAGCTGAAGAGAGAGGCTGACAATTTAAAGGACAGA  
4. AACGAGGAGCTCAAGCGAGAAGCTGATACGCTGAAGGAC

Could this have functional implications?

The resulting proteins from sequences 1 and 2 are the same. However, sequences 1 and 2 differ slightly, and hence the same proteins are formed with different codons as building blocks. For example, both sequences have one alanine amino acid, but it is produced by the GCT codon in sequence 1 and the GCC codon in sequence 2.

Now use the Needleman Wunsch algorithm to compare sequence 1 to the other sequences. Use the scoring: match +2, mismatch -3, indel -4. Also perform the comparison with sequence 4 on paper, using the first three codons only. Comparing the scores, what would you conclude about the relatedness of the species? Is this result consistent with know phylogenetic relatedness?

Different codons can produce the same amino acids, but they different codons have different frequencies between species, as I looked at a little in question 4.

1 with 2 48

with 3 -27

with 4 -17

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Papio anubis potassium voltage-gated channel subfamily B member 1 (KCNB1),

Gorilla gorilla gorilla potassium voltage-gated channel subfamily B member 1 (KCNB1)

Callithrix jacchus potassium voltage-gated channel subfamily B member 1-like

Cebus capucinus imitator potassium voltage-gated channel subfamily B member 1 (KCNB1)

Homo sapiens potassium voltage-gated channel subfamily B member 1 (KCNB1)

Pan troglodytes potassium voltage-gated channel subfamily B member 1 (KCNB1)

Macaca fascicularis potassium channel, voltage gated Shab related subfamily B, member 1 (KCNB1)

Pan paniscus potassium channel, voltage gated Shab related subfamily B, member 1 (KCNB1)

Nomascus leucogenys potassium channel, voltage gated Shab related subfamily B, member 1 (KCNB1)

Cercocebus atys potassium channel, voltage gated Shab related subfamily B, member 1 (KCNB1)

Macaca nemestrina potassium channel, voltage gated Shab related subfamily B, member 1 (KCNB1)

Mandrillus leucophaeus potassium channel, voltage gated Shab related subfamily B, member 1 (KCNB1)

Chlorocebus sabaeus potassium voltage-gated channel, Shab-related subfamily, Macaca mulatta potassium voltage-gated channel subfamily B member 1 (KCNB1)

